LIGO Public Data

Lessons learned

https://losc.ligo.org/

Jonah Kanner

LIGO Lab, Caltech

June 12, 2017

G1701082

- Making data public is easy, making it usable is hard
  - Data quality, segments, meta-data, documentation, tutorials, examples, software, spacecraft state information

- Some users want to be like you …
  - Tutorials very popular, and used for student training, classroom activities, etc
  - Popular for training next generation of GW scientists
  - Some will see examples as "right" way to do things
  - Important to give notes about common pit-falls
    - What are the limitations of your data? What pre-processing is required?

- … but some do **not**.
  - Some will use own software, not yours. Maybe not what you expect
  - Common data format important. We are routinely asked for ASCII or CSV
  - 95% of computers run Windows. What software tools will they use?
  - Excel is popular.
  - Things that run in the browser are good (we like IPython notebooks)
  - Audio files, pre-made plots, pre-processed data are all popular
    - But may be misused
  - Artists / amateur scientists / young students
  - Visual / video instructions are good

- E-mail list gets used
  - For us, a ticketing system has really helped
  - People will ask for projects / mentoring
    - To what extent will you support this?
  - Questions not limited to technical
    - Where can you refer EPO / general questions?
  - Will you advertise projects done with your data?

- Managing public releases can be a challenge
  - Need to develop web site, but not release secrets
  - Google finds anything public, and never forgets
  - DOIs, URL names, static files, all need careful management

- Important to keep stats – we get asked routinely
  - Number of downloads, Google analytics, number of citations, etc.
  - We are struggling to track publications

- How will data be organized and “discovered”?
  - Do you need a database to describe your data set?

Thank you

Extra slides

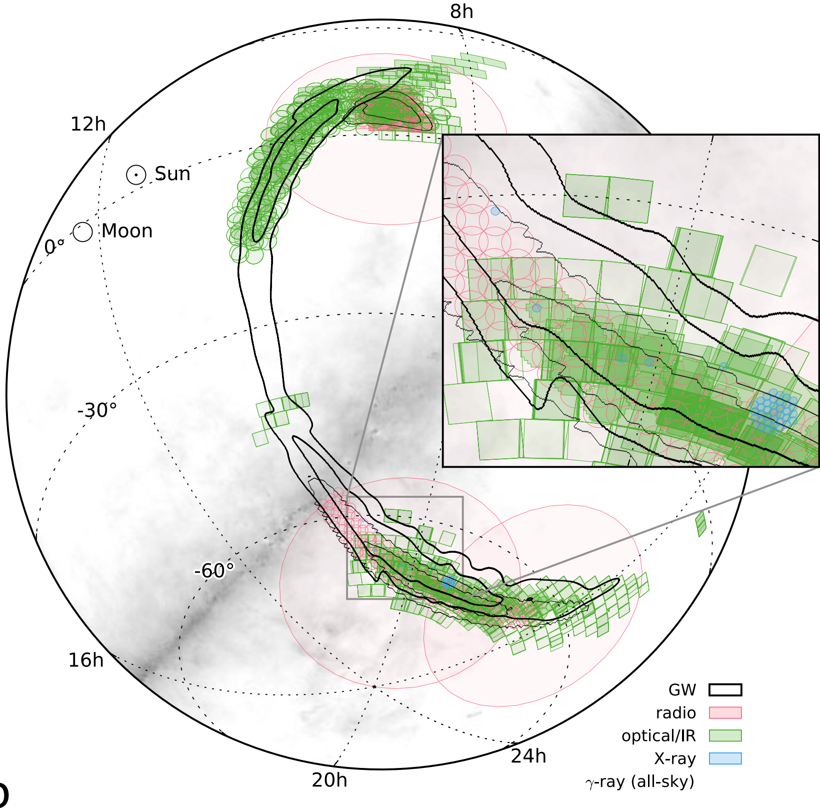
What is LIGO data?

- 1 gravitational wave sensitive channel per detector
  - Sampled at 16kHz (30 kB / s) or 1-ish TB per year
- Thousands of "auxiliary channels"
  - Sample rates vary
  - 25 MB/s or 1-ish PB per year
- Stored in international standard file format
  - IGWD Frames
  - Frame file may contain many channels
  - Libraries available to work with frames:
    - FrameCPP, framelib, gwpy, LAL, …
  - Also use HDF5 for public data releases
- Low-latency "triggers" as GCN alerts

Low-latency Triggers

- Allow EM follow-up of LIGO transients
  - Follow model of gamma-ray burst community
- Include key properties of event:
  - time, significance, source position, source type, …
- Available in about 5 minutes
  - Distributed after human validation
- Enthusiastic response
  - MOUs w/ 80 astronomy collaborations
  - Around 25 teams observed in response to GW150914

GW150914 Follow-up



ApJ Letters, Volume 826, Number 1

LIGO Data Management Plan

- Overview of LIGO data preservation and access:
  - http://dcc.ligo.org/LIGO-M1000066/ (Included in pre-review documents)
  - Updated annually

- For LSC scientists:
  - "Bulk" data is copied to several LIGO computing centers (LIGO Data Replicator)
  - "Stream-based" data is available through network data servers [NDS2]
  - Provide authenticated data access on shared resources
    - OSG, XSEDE, Blue Waters … [CernVM FS]
  - Data is preserved on a tape drive archive

- For the public:
  - The LIGO Open Science Center provides access to calibrated h(t)
  - Meta-data, Documentation, Tutorials, Software tools
  - Plan details timeline for data releases

Why public data?

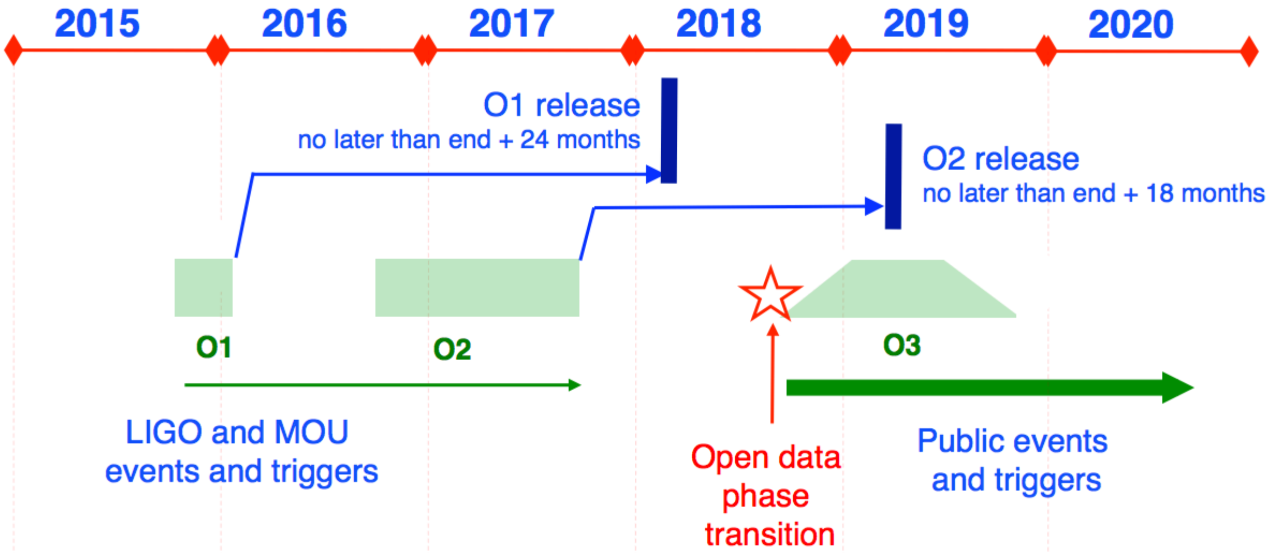
- Enable broadest participation in LIGO research
  - Better science
  - Wider research community
    - LIGO scientists, astrophysics, theory, NR, astronomy …
  - Amateur scientists
  - Student training, teachers, workshops, and EPO

- Broad national movement toward open data
  - E.g. OMB Open Data Memorandum, Project Open data …
  - Requirement from the NSF

Two Phases for Open Data

- Phase 1: Discovery Phase
  - 1.1 hours (4096 s) of data around all discoveries … and other interesting times (e.g. GRBs)

- Phase 2: Observational Phase
  - Release ALL strain data in 6 month blocks … after 18 month proprietary period
  - Public low-latency alerts for transients



Open Data: Status

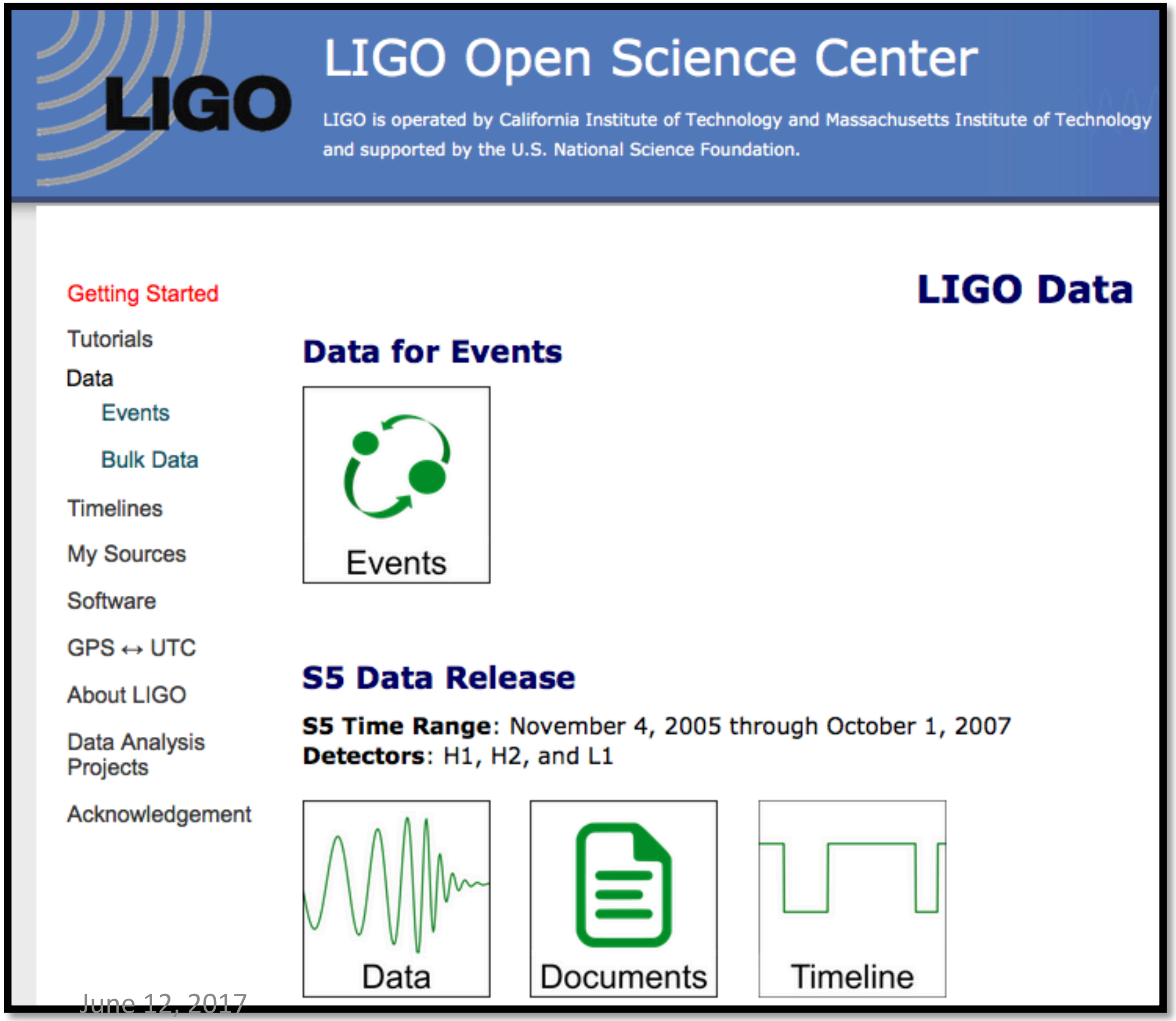
- In discovery phase:
  - ✓ Released data around 3 BBH discoveries!

    … plus data around candidate event LVT151012
  - ✓ (Added) Released Initial LIGO strain data

    3 years of S5 and S6 data (2005-2010)
- Begin open data era at beginning of O3
  - – Achieved milestone of "plentiful detections"
  - – O1 will be released w/ 2 year lag
  - – For O2/O3, shorten proprietary period to 18 months
- Draft O1 data set under review

Public Data Access: LOSC

https://losc.ligo.org/



LIGO Open Science Center

LIGO is operated by California Institute of Technology and Massachusetts Institute of Technology and supported by the U.S. National Science Foundation.

LIGO Data

Data for Events

S5 Data Release

S5 Time Range: November 4, 2005 through October 1, 2007
Detectors: H1, H2, and L1

Easy point & click downloads of calibrated strain data

Includes:

- Data Discovery
- Documentation
- Examples
- Data Quality
- Segments

LOSC: “Bulk” data download

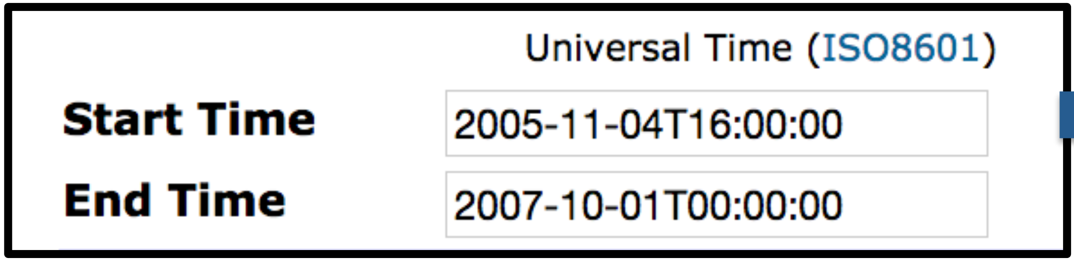
Simple query by start/stop time:

→Returns list of data files to download

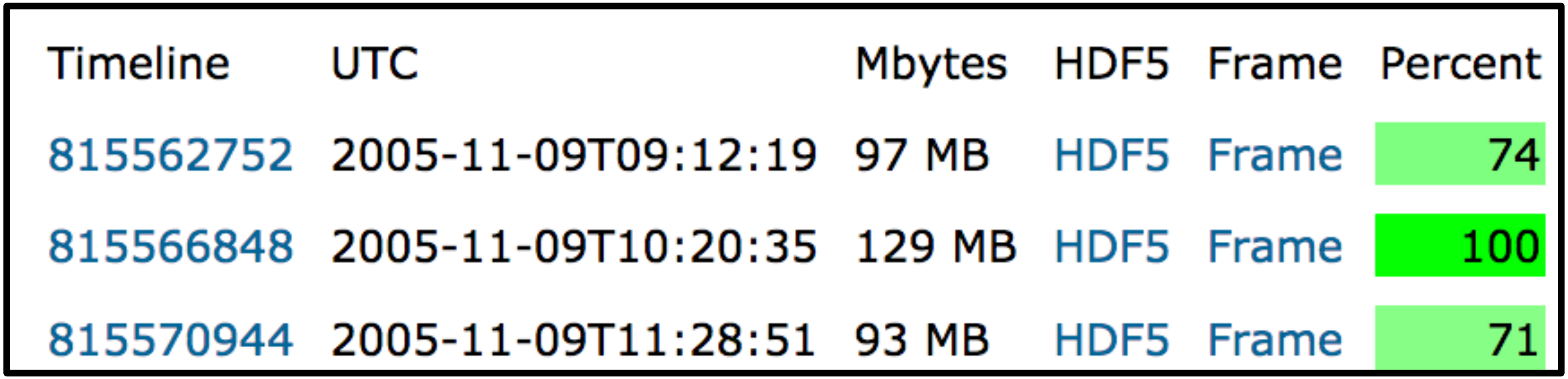
→Choice of HDF5 or Frame

→Python API to read both formats (readligo.py)

→Predictable URL’s and JSON file lists
for automated downloads



	Universal Time (ISO8601)
Start Time	2005-11-04T16:00:00
End Time	2007-10-01T00:00:00



Timeline	UTC	Mbytes	HDF5	Frame	Percent
815562752	2005-11-09T09:12:19	97 MB	HDF5	Frame	74
815566848	2005-11-09T10:20:35	129 MB	HDF5	Frame	100
815570944	2005-11-09T11:28:51	93 MB	HDF5	Frame	71

LOSC: Event pages

Data release around times of LIGO discoveries

→4096 seconds of strain + data quality

→GWF / HDF5 / ASCII

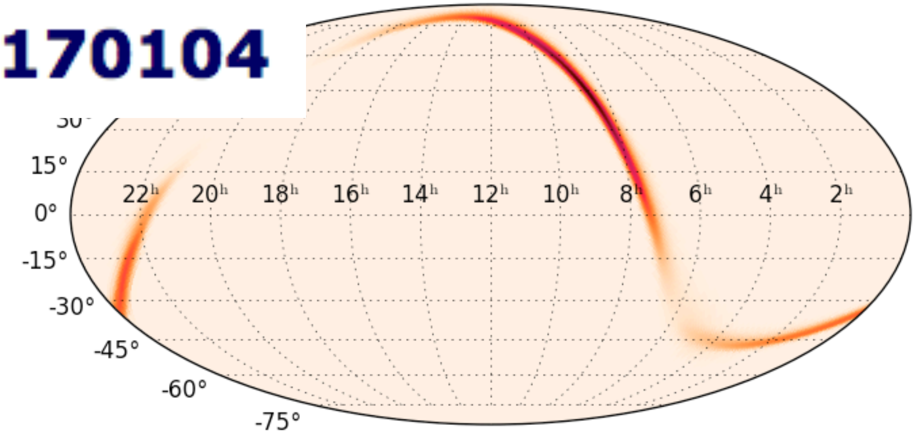
→Skymaps

→Parameters and best fit waveform

→Documentation & tutorials

Data release for event GW170104

https://losc.ligo.org/events/



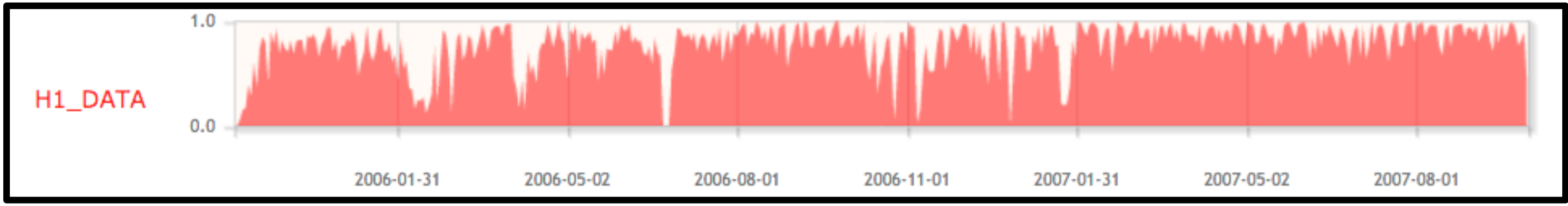
Data Quality / Segments (Timeline)

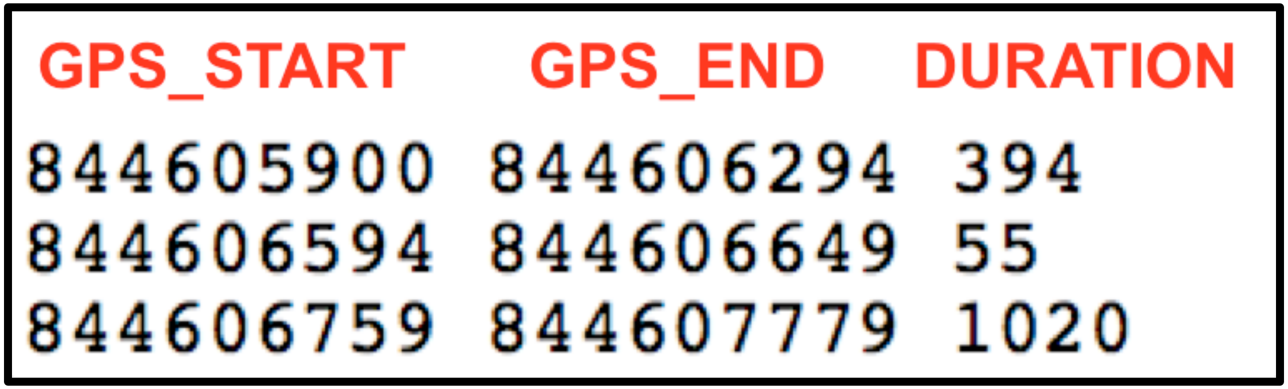
Provide 1 Hz data quality channels

→ Times data is available

→ CAT 1/2/3, convention used by working groups

→ Available in files, segments lists, interactive plots



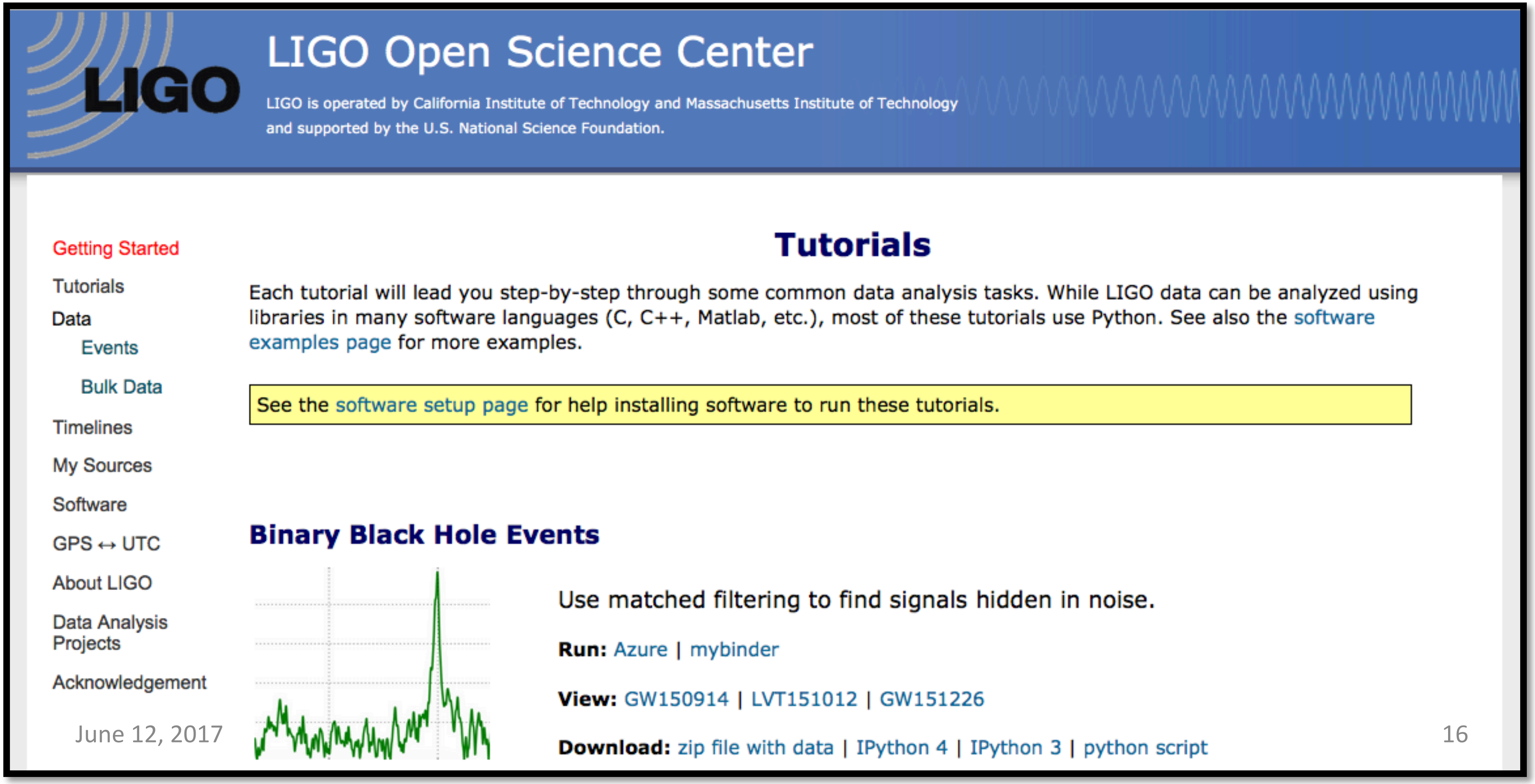


GPS_START	GPS_END	DURATION
844605900	844606294	394
844606594	844606649	55
844606759	844607779	1020

Tutorials

https://losc.ligo.org/tutorials/

Examples use python to load data, make plots, find signals



LIGO Open Science Center

LIGO is operated by California Institute of Technology and Massachusetts Institute of Technology and supported by the U.S. National Science Foundation.

Tutorials

Each tutorial will lead you step-by-step through some common data analysis tasks. While LIGO data can be analyzed using libraries in many software languages (C, C++, Matlab, etc.), most of these tutorials use Python. See also the software examples page for more examples.

See the software setup page for help installing software to run these tutorials.

Binary Black Hole Events

Use matched filtering to find signals hidden in noise.

Run: Azure | mybinder

View: GW150914 | LVT151012 | GW151226

Download: zip file with data | IPython 4 | IPython 3 | python script

Tutorials

https://losc.ligo.org/tutorials/

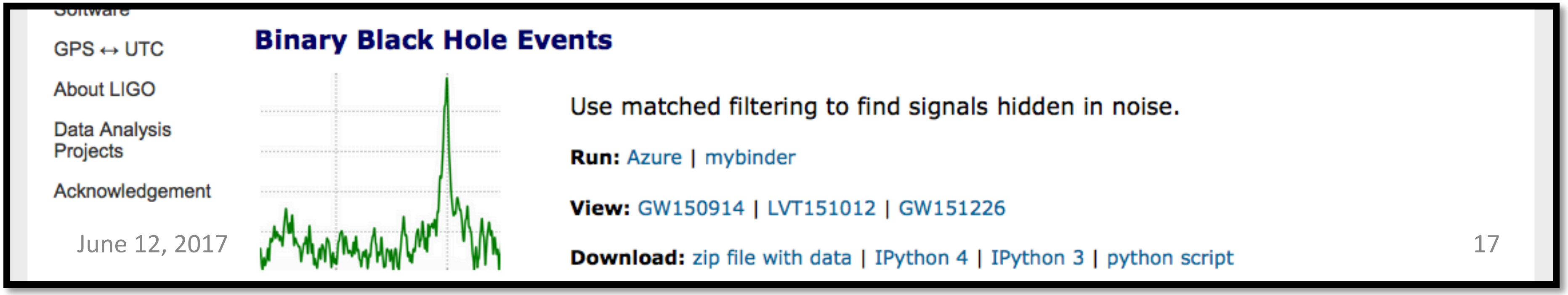
Three ways to access tutorials: Run, View, or Download

Run: Run tutorials in your browser with “binder” or “Microsoft Azure” iPython Notebooks

- Binder provides instant access, no log-in
- Microsoft Azure provides log-in feature to save work, create, & share new notebooks

View: See the tutorial as an HTML web page

Download: Download the code and run it on your own computer



Software

GPS ↔ UTC

About LIGO

Data Analysis Projects

Acknowledgement

Binary Black Hole Events

Use matched filtering to find signals hidden in noise.

Run: Azure | mybinder

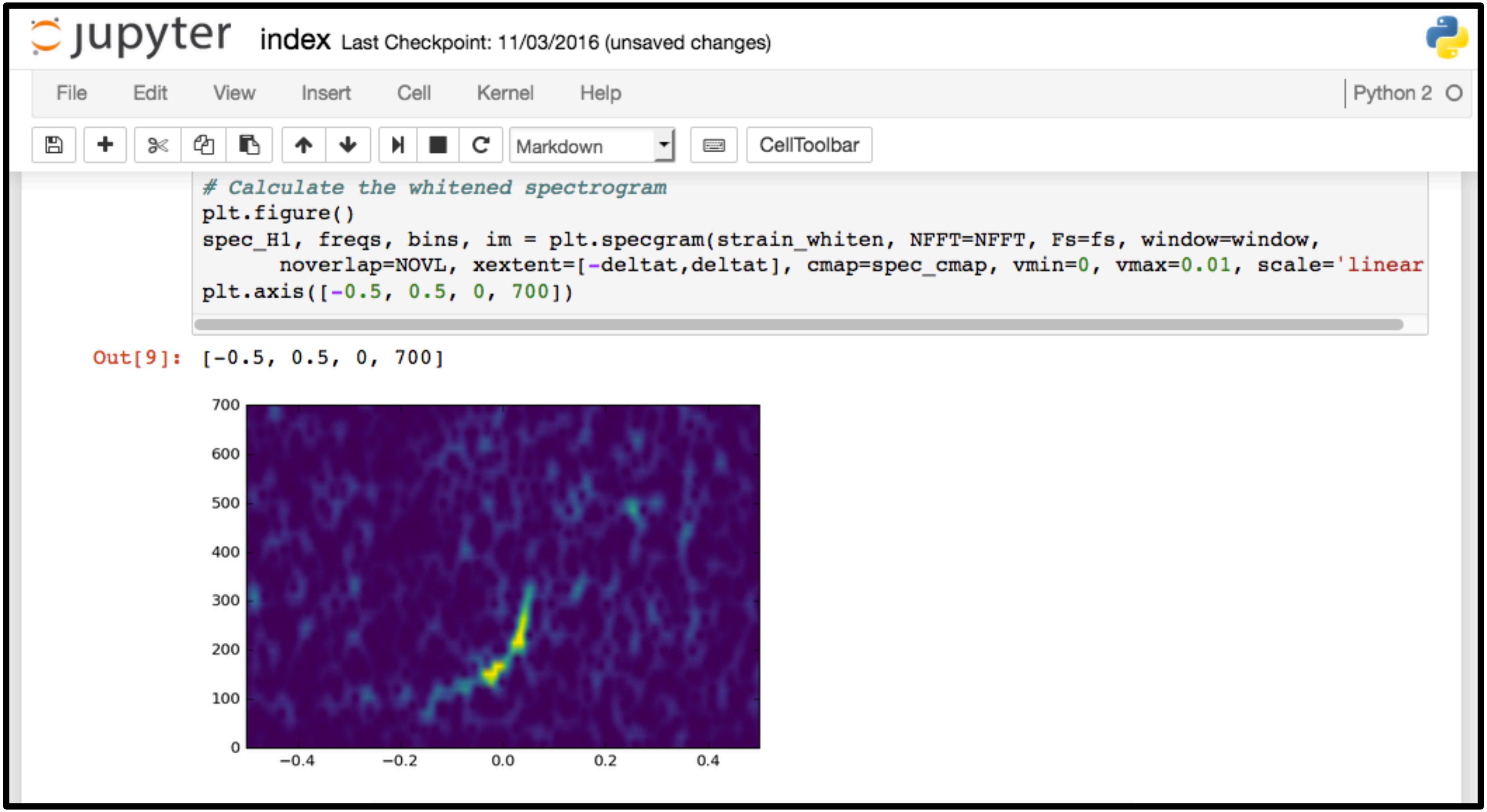
View: GW150914 | LVT151012 | GW151226

Download: zip file with data | IPython 4 | IPython 3 | python script

LOSC Tutorials

Users interact with LIGO data through the web browser

…. No software installation required



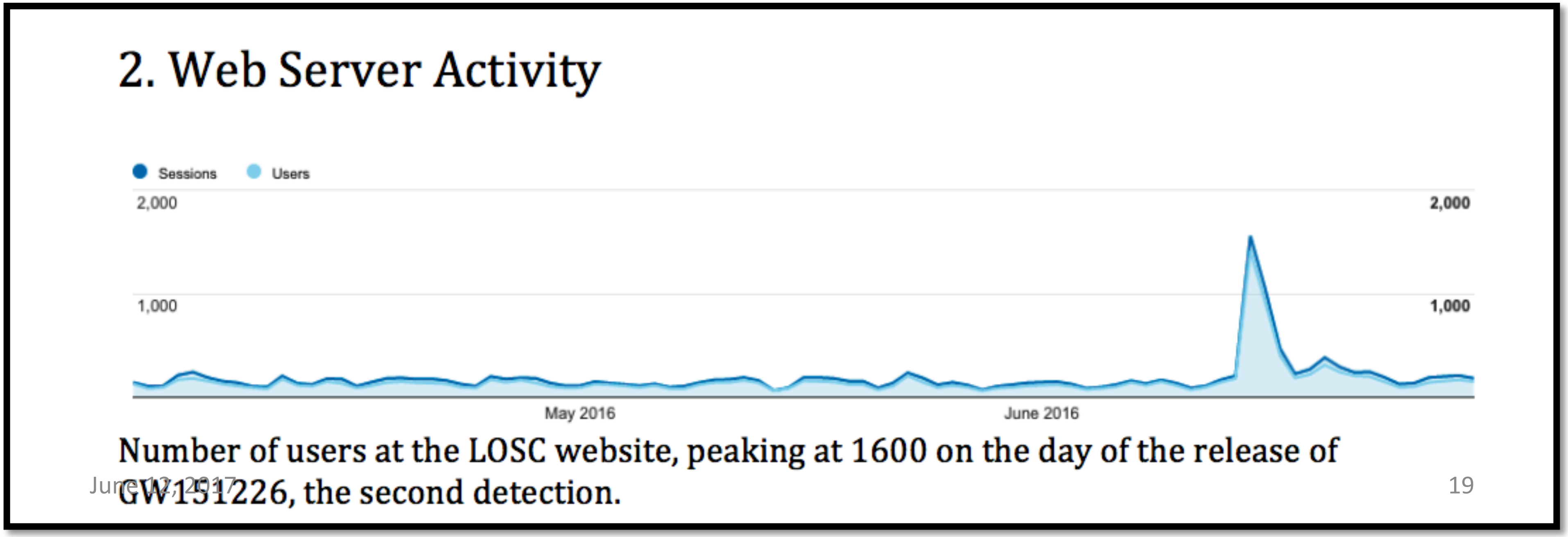
```
# Calculate the whitened spectrogram
plt.figure()
spec_H1, freqs, bins, im = plt.specgram(strain_whiten, NFFT=NFFT, Fs=fs, window=window,
      noverlap=NOVL, xextent=[-deltat,deltat], cmap=spec_cmap, vmin=0, vmax=0.01, scale='linear
plt.axis([-0.5, 0.5, 0, 700])
```

```
Out[9]: [-0.5, 0.5, 0, 700]
```

LOSC Usage

Track web site usage through Google Analytics

- Example stats in LOSC Q2 report: https://dcc.ligo.org/LIGO-P1600244
- Typically about 100 users per day
  - 50% new / 50% returning
  - Over 26,000 users over the past year
  - Typically several hundred data file downloads per day
  - Visitors from all 50 states and all around the world
  - GW150914 page and tutorials are most popular pages



2. Web Server Activity

Number of users at the LOSC website, peaking at 1600 on the day of the release of GW151226, the second detection.

LOSC Usage

- Used for training of young GW scientists
  - Summer schools, new grad students, KAGRA, IndIGO, conferences
  - Tutorials are popular
- Used for student projects
  - High school, undergrads, science fair, art projects, citizen scientists
- Scientific publications
  - Aware of a handful
  - Looking for a good tool to track this
    - Already ask authors to acknowledge LOSC and NSF
- Classroom activities
  - Lab activities, teacher training, text book problems

- See https://losc.ligo.org/projects/